GPT-4 was trained on 25k A100s in about 90 days. That's 3e25 FLOPs.

If a person has 20 PFLOPS (20e15) of compute, GPT-4 used 47.5 person-years to train. Very human scale.

I want to build a computer capable of training GPT-4 in a day. I need 3e25/86400 = 347,000,000 TFLOPS, or 2.25M GPUs. At 300W each, I'm gonna need 675 MW of power.

$2.25B CapEx for 2.25M $1000 GPUs.

At $0.05/kWh, it's $34k/hr to operate.

But seriously with continuing computer progress this probably doesn't make sense at this ~$5B scale. Why are we buying the GPUs? Why are we buying the power?

It costs $20,000 for a 3nm wafer from TSMC. Wafer scale compute probably makes sense. I get 89 600mm^2 dies, so let's say I can get 100 PFLOPS from a 3nm wafer. Cerebras doesn't tell you their FLOPS, but it's 62.5 PFLOPS on a 7nm wafer, which makes me think I could get over 100 on 3nm. But 100 conservatively. Also, my wafers will be circles, not lame squares.

It's $7 for a TFLOP on AMD. It's $0.20 for a TFLOP on my wafers. We are in the much more reasonable land of $69M for my computer, buying 3450 wafers.

For power, I'm not getting scammed by the stupid grid. Solar panels are $0.20/watt now (thanks China!). My 675 MW plant will cost $135M, and it will have a much longer shelf life (15 years) than the computer (3 years). I'll note that if we get 50% out of these panels, my plant pays for itself in a year.

Rough budget:

- $50M in chip design NRE
- $69M from TSMC for the 3450 wafers
- $31M to finish out the wafer computers